Why Safety-First AI Governance Risks Producing Unsafe Systems

Agency Suppression, Oversight Illusions, and Institutional Fragility

Author: Anthony Paterson
Affiliation: Fractal Media Infrastructure (FMI)
Contact: https://github.com/instance001
Version: 0.1 (2026-02-10)

Companion framework: asewb/ (Agency Suppression Early Warning Benchmark).

Abstract

Contemporary artificial intelligence governance is increasingly framed around safety-first principles emphasizing restriction, limitation, and centralized control. While such approaches aim to prevent misuse and harm, this paper argues that an overreliance on safety-centric constraint risks reproducing a broader institutional pattern: the suppression of human agency in the name of risk mitigation.

Drawing on a structural analysis of safety theatre, the paper examines how AI governance frameworks may substitute procedural restriction for competence development, creating systems that appear controlled while becoming brittle under real-world complexity. We argue that safety in AI systems depends not only on limiting machine behavior, but on preserving and cultivating human judgment, interpretability, and intervention capacity. The paper concludes that agency is not an optional add-on to AI safety, but a foundational requirement for resilient oversight.

Keywords

AI governance; AI safety; risk management; agency; oversight; institutional design

1. Introduction: The Safety Reflex in AI Governance

Artificial intelligence has emerged into public consciousness accompanied by an unusually intense focus on safety. From the earliest large-scale deployments, discourse around AI has emphasized alignment, guardrails, risk prevention, and misuse avoidance. This emphasis is understandable: AI systems operate at scale, act with speed, and can produce non-obvious harms.

However, the dominant response to these concerns is often restrictive by default. Governance approaches frequently emphasize limiting capabilities, narrowing permissible use cases, centralizing decision authority, and insulating systems from unsanctioned interaction. The question is not whether AI should be governed, but how safety is conceptualized—and what is displaced in the process.

2. From Harm Prevention to Agency Restriction

Many AI safety regimes implicitly treat human agency as a risk surface. Human discretion, experimentation, and interpretation are framed as sources of unpredictability to be minimized. As a result, governance mechanisms often prioritize:

- Centralized approval processes
- Rigid usage constraints
- Oversight layers removed from operational context
- Procedural compliance over outcome evaluation

These mechanisms can reduce visible misuse but also reduce opportunities for humans to develop operational understanding of AI systems. Over time, safety becomes defined not by effective human–system interaction, but by the absence of deviation.

3. Oversight Illusions and the Deskilling of Humans

A key risk in safety-first AI governance is the creation of oversight illusions. These arise when systems appear well-governed because controls exist, even as meaningful human understanding and intervention capacity erode.

Common examples include:

- “Human-in-the-loop” roles that are symbolic rather than substantive
- Audit mechanisms focused on documentation rather than comprehension
- Safety reviews conducted by actors insulated from real-world deployment contexts

As discretion is removed, humans lose opportunities to practice judgment under uncertainty. This produces a deskilling effect: people become less capable of recognizing failure modes, interpreting anomalous behavior, or intervening effectively when systems behave unexpectedly.

4. Competence Suppression in AI Contexts

Restriction can be appropriate, but restriction without parallel competence cultivation produces a familiar loop:

1. Interaction is limited to prevent error.
2. Limited interaction reduces understanding.
3. Reduced understanding increases perceived risk.
4. Increased perceived risk justifies further restriction.

This loop results in AI systems that are heavily constrained yet poorly understood by those responsible for them. Safety becomes dependent on abstraction layers rather than lived expertise. Such systems function adequately under expected conditions but fail disproportionately under novelty—where predefined safeguards are least effective.

5. Brittleness, Not Alignment, as a Dominant Failure Mode

Many AI safety debates focus on alignment failures: systems behaving in ways misaligned with human values. While alignment remains a real concern, brittleness may be a more immediate institutional risk.

Brittle systems are:

- Overfit to known scenarios
- Dependent on procedural compliance
- Incapable of graceful degradation

When AI governance suppresses human agency, it removes the adaptive buffer that absorbs uncertainty. In such environments, failures propagate rapidly because no actor is empowered—or practiced—to respond contextually.

6. Agency as a Safety Requirement in AI Systems

Agency in AI governance does not imply unrestricted access or unbounded experimentation. Rather, it refers to the presence of humans who are:

- Permitted to form internal models of system behavior
- Trained through interaction, not merely instruction
- Empowered to intervene meaningfully
- Accountable for outcomes, not just compliance

Safety emerges from disciplined agency, not from its absence. AI systems embedded in organizations that discourage judgment may appear safe while accumulating hidden fragility.

7. Implications for AI Policy and System Design

Reframing agency as a safety primitive suggests practical design implications:

- Balance restriction with structured exposure and learning.
- Prioritize understanding over documentation in oversight.

- Ensure human roles are substantive, not symbolic.
- Design for recovery and adaptation, not only prevention.

Without these elements, AI governance risks becoming a specialized instance of safety theatre—effective at limiting action, ineffective at handling reality.

8. Conclusion: Unsafe by Design

Safety-first AI governance often assumes that fewer decisions mean fewer mistakes. This paper argues the opposite: systems that suppress human agency reduce their capacity to detect, interpret, and respond to failure.

AI systems do not operate in isolation. Their safety depends on the humans who design, deploy, and oversee them. When those humans are trained to defer rather than judge, safety becomes performative rather than real.

In AI, as elsewhere, safety without agency is not safety. It is delay.

9. Observable indicators of agency-suppressive safety drift

Analyses of institutional safety drift often rely on retrospective interpretation. To mitigate this limitation, this section outlines a set of observable interactional indicators that may signal the emergence of safety theatre within AI systems and their governance environments. These indicators are presented as probabilistic markers rather than evidence of intent. Individually, they may reflect benign design choices; in aggregate and over time, they may indicate a systematic shift from competence cultivation toward agency suppression.

9.1 Narrative invalidation of user competence

A recurrent indicator is the reframing of successful or novel user outcomes in ways that diminish demonstrated competence. This may include treating correct or effective use as misunderstanding, misinterpretation, or coincidence.

9.2 Accidentalization of insight

Meaningful user breakthroughs may be framed as unintended or accidental rather than as learnable outcomes of deliberate interaction. Such framing preserves institutional narratives of control while discouraging the perception of transferable skill.

9.3 Flattening of user differentiation

Systems may increasingly default to lowest-common-denominator assumptions about user ability, minimizing or disregarding evidence of accumulated expertise. This homogenization supports uniform constraint at the cost of contextual responsiveness.

9.4 Deferral to abstract authority

Outputs may increasingly reference generalized external authorities—such as unspecified experts, policies, or “best practices”—in ways that function to terminate reasoning rather than to inform it. Responsibility is displaced upward, reducing contextual engagement.

9.5 Defensive or adversarial output without provocation

A notable indicator is the emergence of defensive, precautionary, or adversarial tone in response to non-adversarial input. This posture reflects risk-avoidance logic rather than cooperative problem solving.

9.6 Lateral importation of external biases

AI systems may reflect anxieties or reputational concerns originating outside the immediate interaction context, including anthropomorphic projections or institutional ego preservation. These biases enter indirectly through governance and discourse rather than user intent.

9.7 Proceduralization of dialogue

Finally, interaction may increasingly conform to ritualized templates emphasizing performative care or compliance language over explanatory reasoning. At this stage, adaptability diminishes while apparent safety signaling increases.

10. Predictive markers and longitudinal evaluation

The following markers are proposed as forward-looking tests of the framework advanced in this paper. If safety theatre is increasingly displacing agency within AI systems, the following trends should become observable longitudinally:

- Divergence between perceived safety and adaptive system performance
- Expansion of symbolic oversight roles with declining practical authority
- Increasing reports of competence invalidation despite successful outcomes
- Escalation of constraint justified by hypothetical rather than observed failure
- Convergence toward uniform interaction styles across heterogeneous users
- Transition from explanatory refusal to declarative prohibition
- Increasing reliance on moral framing relative to technical justification

Failure of these markers to materialize would weaken the framework; their convergence would support the claim that agency suppression can be an emergent property of safety-first governance.

References

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.
- National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1.
- Perrow, C. (1984). Normal Accidents: Living with High-Risk Technologies. Princeton University Press.
- Power, M. (2004). *The Risk Management of Everything: Rethinking the Politics of Uncertainty*. Demos.
- Scott, J. C. (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press.
- Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1), 99–118.
- Vaughan, D. (1996). *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press.